On the underfitting and overfitting sets of models chosen by order selection criteria

Xavier Guyon, Jian-Feng Yao

Centre de Recherche

SAMOS

Statistique Appliquée &
Modélisation Stochastique

On the underfitting and overfitting sets of models chosen by order selection criteria

X. GUYON et J.F. YAO

Prépublication du SAMOS N° 71 (Version Révisée)

Mai 1998

On the underfitting and overfitting sets of models chosen by order selection criteria

Xavier GUYON and Jian-feng YAO

SAMOS - Université Paris 1*

* e-mail: guyon[yao]@univ-paris1.fr

Running head : Underfitting and overfitting sets of models

Mail address for correspondence : J. F. YAO , SAMOS, Université Paris 1, 90 rue de Tolbiac, 75634 Paris Cedex 13, France

Abstract

For a general class of order selection criteria, we establish analytic and non asymptotic evaluations of both the underfitting and overfitting sets of selected models. These evaluations are further specified in various situations including regressions and autoregressions with finite or infinite variances. We also show how upper bounds for the misfitting probabilities and hence conditions ensuring the weak consistency can be derived from the given evaluations. Moreover, it is demonstrated how these evaluations, combined with a law of the iterated logarithm for some relevant statistic, can provide conditions ensuring the strong consistency of the model selection criterion used.

Key Words. Model selection ; AIC ; BIC ; underfitting and overfitting ; weak consistency ; strong consistency ; regressions and autoregressions ; Markov fields ; stable law

AMS Classification Numbers : Primary 62 F 12 ; Secondary 62 M 10, 62 M 40

1 Introduction and notations

Let $X = (X_1, X_2, ...)$ be a sequence of observations generated by a semiparametric distribution $\mathbb{P}_{\theta_0}$ where θ_0 is the finite dimensional part of the true model and belongs to Θ, a subset of $\mathbb{R}^m$. The integer m should be thought as an upper-bound for the dimension of the parametric part. The goal of model selection is to estimate the true model.

We call any subset P of $M = \{1, ..., m\}$ a *submodel* and identify P with the parameter subspace $\Theta_P = \Theta \cap \{\ \theta_i = 0 \text{ for } i \notin P\}$. The cardinality of P, denoted by p, is the dimension of Θ_P. Thus M corresponds to the *full model*. Our purpose is to find the true model $P_0 = \{i : \theta_{0,i} \neq 0\}$, that is the locations of non-null coordinates of θ_0.

Given n observations $X(n) = \{X_1, X_2, ..., X_n\}$, a general model selection criterion [1] consists of minimizing in P a penalized pseudo-likelihood (PL) or objective function $U_n(\theta) = U_n(\theta, X(n))$: first we estimate θ_0 in P by

$$\widehat{\theta}_P = \text{Arg} \min_{\theta \in \Theta_P} U_n(\theta) \ . \tag{1}$$

Let $(c_n)_{n \geq 0}$ be some sequence of positive numbers (*penalization rate*). We estimate P_0 by

$$\widehat{P}_n = \text{Arg} \min_{P \subseteq M} \left\{ U_n\left(\widehat{\theta}_P\right) + \frac{c_n}{n} p \right\} \ . \tag{2}$$

For instance with $U_n = -\frac{2}{n} \log(\text{Likelihood})$, we obtain the AIC for the constant rate $c_n = 2$, while the rate $c_n = \log n$ yields the well-known BIC criterion.

Therefore, a submodel P will be preferred to the true model P_0 if and only if:

$$\Delta_n(P, P_0) := U_n(\widehat{\theta}_P) - U_n(\widehat{\theta}_{P_0}) \leq \frac{c_n}{n}(p_0 - p). \tag{3}$$

The *underfitting set* M_n^- and the *overfitting set* M_n^+ are respectively

$$M_n^- = \left\{ \ \widehat{P}_n \not\supseteq P_0 \ \right\}, \qquad M_n^+ = \left\{ \widehat{P}_n \supsetneq P_0 \right\} \tag{4}$$

The purpose of this paper is to provide an accurate evaluation of these two misfitting sets in a unified and general set-up. Our main assumption is that the PL process $U_n(\theta)$ could be *factorised* as $U_n(\theta) = U(\theta, T_n)$, where U is a known deterministic function and T_n some sample statistic. The remaining assumptions on model identifiability or its smoothness are more standard.

Our main results (Theorems 1 and 2) give, *for a fixed sample size* n, the evaluations of M_n^- and M_n^+. These evaluations are not asymptotic, hence they can be used for small or moderate sample sizes n. Another important feature is that these evaluations are *analytic:* by this we mean that they are derived without using any stochastic properties of the models. Actually, they only depend on the smoothness of the map $(\theta, \gamma) \mapsto U(\theta, \gamma)$.

Furthermore, these evaluations shed a new light on the known asymmetry between the two misfitting sets M_n^- and M_n^+. For example, we can easily see from these evaluations how the overfitting set M_n^+ depends on the penalization rate c_n much more than the underfitting set M_n^-.

As an important application, we will use these evaluations to derive upper bounds for the misfitting probabilities $\mathbb{P}(M_n^-)$ and $\mathbb{P}(M_n^+)$. Consequently, sufficient conditions on the rate c_n will be given to ensure the weak consistency. On the other hand, if the

almost sure convergence rate of the statistic T_n can be estimated through e.g. a law of the iterated logarithm, strong consistency of the selection criterion can be derived in a straightforward way.

Such a penalization criterion for model selection was first introduced by Akaike [1]. Since the literature on the subject is huge, we just mention some references related to the applications developed in this paper. Strong consistency is established for various linear models by [12, 13], [24], [9], [29], [20], [32], [15], [22, 23] and [27].

Probability estimates of the misfitting sets M_n^- and M_n^+ has been much less studied. In the case of an AR process, [28] obtained for the AIC criterion an exact evaluation of the overfitting probability $\mathbb{P}(M_n^+)$. In the same context, Bai and al. [2] proposed an upper bound for $\mathbb{P}(M_n^+)$. Their approach has largely inspired our work. Other related results can be found in [4] for convolution models, in [3] for a log-linear models and in [31] for regression models.

The paper is organized as follows. In Section 2, we establish the main theorems. In Section 3, we apply these results in various situations: regression models with least squares estimation, Whittle's PL for an AR process or a CAR Markov field on $\mathbb{Z}^d$, categorical data models with maximum-likelihood estimation, and Markov fields with Besag's PL estimation. For these models, we establish in Section 4 upper bounds for misfitting probabilities $\mathbb{P}(M_n^-)$ and $\mathbb{P}(M_n^+)$. Weak consistency of the selection criterion is derived under suitable condition on c_n. Finally, we show in Section 5 how our evaluations, combined with a law of the iterated logarithm for the statistic T_n, can be used to address the strong consistency of the model selection procedure.

2 Evaluation of the misfitting sets M_n^- and M_n^+

Let us first introduce some notations. For any scalar map ϕ depending on some vector variables, say c and d, we shall denote its derivatives by $\phi_c^{(1)} = (\partial\phi/\partial c_j)$ and $\phi_{c,d}^{(2)} = (\partial^2\phi/\partial c_i \partial d_j)$. The maximum and minimum of two real numbers u, v are denoted by $u \vee v$ and $u \wedge v$, respectively. The norm $\|\cdot\|$ and the inner product $\langle \cdot, \cdot \rangle$ are Euclidean. For a linear map A from $\mathbb{R}^p$ to $\mathbb{R}^q$, we use the operator norm $\|A\| = \sup\{\|Au\| : \|u\| = 1\}$. The open ball with center x and radius r is denoted by $B(x\,;\,r)$ and the transpose of a matrix A by A'.

Let n be some fixed positive integer. In the following assumptions, **(C.2)** and **(C.3)** are defined with respect to some fixed point $\gamma_0 \in F \subseteq \mathbb{R}^k$.

(C1) Factorization. For an open parameter space $\Theta \subset \mathbb{R}^m$ and some integer k, there is some statistic $T_n = T_n(X_1, X_2, ..., X_n) \in F \subseteq \mathbb{R}^k$ and a *continuous* map $U : \Theta \times F \to \mathbb{R}$ such that $U_n(\theta) = U(\theta, T_n)$.

(C2) Identifiability. (i) For all $\theta \in \Theta$, $U(\theta, \gamma_0) \geq U(\theta_0, \gamma_0)$. (ii) If $U(\theta, \gamma_0) = U(\theta_0, \gamma_0)$, then $\{i : \theta_i \neq 0\} \supseteq P_0$.

(C3) Smoothness. There is some open ball $V := B(\gamma_0\,;\,R_1)$ in F such that, for all $P \subseteq M$ and $\gamma \in V$

(i) the map $\theta_P \mapsto U(\theta_P, \gamma)$ from Θ_P to $\mathbb{R}$ has a minimum $\theta_P(\gamma) \in \Theta_P$;

(ii) these minima can be selected such that the map $\gamma \mapsto \theta_P(\gamma)$ is continuous on V.

The fixed point γ_0 corresponds to a central value of T_n and will be specified in examples below. Clearly, the identifiability assumption **(C2)** is fulfilled if θ_0 is the unique global minimum of the map $\theta \mapsto U(\theta, \gamma_0)$ on Θ. This will happen for applications carried out in Sections 3 and 4. However, this uniqueness is not necessary. We require instead that P_0 is *minimal*, i.e. that any other model P must contain P_0 if it yields the same minimum (e.g., this can be useful for ARMA models).

Roughly speaking, the smoothness assumption **(C3)** requires that in each submodel P, the PL process $U_n(\theta) = U(\theta, T_n)$ can be *continuously* minimized if the statistic T_n is close to γ_0. Here again, the minimization map $\gamma \mapsto \theta_P(\gamma)$ may not be unique. In particular, the estimator $\hat{\theta}_P = \theta_P(T_n)$ can be any of the possible solutions of (1).

It follows from **(C3)** that for each submodel P, the map $\gamma \mapsto \sigma_P(\gamma) := U[\theta_P(\gamma), \gamma]$ is continuous on V. Since the number of submodels is finite, there is a common modulus of continuity Φ (resp. Ψ) for $\{\sigma_P\}$ of all submodels $P \not\supseteq P_0$ (resp. overmodels $P \supseteq P_0$). More precisely, Φ, Ψ are positive and increasing maps defined on the interval $[0, R_1]$ such that $\lim_{u \to 0_+} \Phi(u) = 0$, $\lim_{u \to 0_+} \Psi(u) = 0$,

$$|\sigma_P(\gamma) - \sigma_P(\gamma_0)| \leq \Phi(||\gamma - \gamma_0||), \text{ for } P \not\supseteq P_0 \text{ and } \gamma \in V \ , \tag{5}$$

and

$$|\sigma_P(\gamma) - \sigma_P(\gamma_0)| \leq \Psi(||\gamma - \gamma_0||), \text{ for } P \supseteq P_0 \text{ and } \gamma \in V \ . \tag{6}$$

Define, for $y \in \mathbb{R}$, the inverse map

$$\Phi^{-1}(y) := \inf\{u : \Phi(u) \geq y\} \quad \text{(by convention, } \inf \emptyset = \infty) \ .$$

We have

(i) $\Phi^{-1}(y) = 0$ for $y \leq 0$, and $\Phi^{-1}(y) = +\infty$ for $y > \Phi(R_1)$.

(ii) $u < \Phi^{-1}(y)$ implies $\Phi(u) < y$, and $y > 0$ implies $\Phi^{-1}(y) > 0$.

The inverse map Ψ^{-1} is defined similarly and satisfies similar properties.

Note that if $P \supseteq P_0$, $\sigma_P(\gamma_0) = U(\theta_0, \gamma_0) = \sigma_{P_0}(\gamma_0)$. On the other hand for any submodel $P \not\supseteq P_0$, by **(C2)**, $\sigma_P(\gamma_0) = \inf_{\Theta_P} U(\theta_P, \gamma_0)$ is strictly greater than $U(\theta_0, \gamma_0)$. Therefore the following constant

$$\Delta_2 := \min \ \{\sigma_P(\gamma_0) - \sigma_{P_0}(\gamma_0) \ : \quad P \not\supseteq P_0\} \tag{7}$$

is positive.

2.1 Main results

Theorem 1 ***(Analytic evaluation of the misfitting sets*** M_n^- ***and*** M_n^+***).*** *Assume that* ***(C1), (C2)*** *and* ***(C3)*** *hold. Then*

(i) The underfitting set M_n^- *satisfies*

$$M_n^- \subseteq \left\{ \|T_n - \gamma_0\| \geq R_1 \wedge \Phi^{-1}\left(\frac{1}{2}[\Delta_2 - m\frac{c_n}{n}]\right) \right\} \ . \tag{8}$$

In particular, setting $\eta_0^- := \Delta_2/(2m)$ *and* $\delta_0^- := R_1 \wedge \Phi^{-1}\left(\frac{1}{4}\Delta_2\right)$*, we have for* $c_n/n \leq \eta_0^-$*,*

$$M_n^- \subseteq \left\{ \|T_n - \gamma_0\| \geq \delta_0^- \right\} \ . \tag{9}$$

(ii) The overfitting set M_n^+ satisfies

$$M_n^+ \subseteq \left\{ \|T_n - \gamma_0\| \geq R_1 \wedge \Psi^{-1}\left(\frac{1}{2}\frac{c_n}{n}\right) \right\} . \tag{10}$$

We now discuss this result while postponing its proof to the end of the section.

Comments.

(i) The evaluations of M_n^- and M_n^+ are by no means asymptotic and hold for each n.

(ii) Often the smoothness **(C3)** holds with large R_1 (even $R_1 = \infty$ as in Section 3.1). In such a case, the term R_1 disappears from the above evaluations.

(iii) Theorem 1 sheds new light on the strong asymmetry between M_n^- and M_n^+. First, the identifiability condition **(C2)** is necessary only for the evaluation of M_n^- (through the constant Δ_2). Second, if c_n/n is small (say $\frac{c_n}{n} < \frac{\Delta_2}{m}$) and $T_n \to \gamma_0$ when $n \to \infty$, we have $M_n^- = \emptyset$ for large n. This is not the case in general for M_n^+ which depends on the relative magnitudes of $||T_n - \gamma_0||$ and c_n/n ([28]).

(iv) The strongest penalization rate is $c_n = \alpha n$ with some $\alpha > 0$. Actually, any positive α is effective to control the overfitting set M_n^+ for large n (assuming again $T_n \to \gamma_0$ as $n \to \infty$). However, the same effectiveness would hold for M_n^- only if the constant α is smaller than $\frac{\Delta_2}{m}$. Since Δ_2 depends on the (unknown) true model P_0, exact evaluation of such an admissible linear rate would be possible only if some additional information about (P_0, θ_0) is available.

To make the evaluation of M_n^+ in (10) more explicit, we need to estimate the modulus of continuity Ψ. In situations where Ψ can be computed in a closed form, application of Theorem 1 is straightforward (see Section 3.1). Otherwise, Theorem 2 below provides a new estimate of M_n^+ by using some second order smoothness of the map U.

(D) Second-order smoothness. There is a radius $r_0 > 0$ (we choose $r_0 < R_1$) such that, for each $P \subseteq M$, the map $U : \Theta_P \times F \to \mathbb{R}$ is twice continuously differentiable on $B_P := B(\mathbf{z}_P\,;\ r_0)$ in $\Theta_P \times F$ with center $\mathbf{z}_P := (\theta_P(\gamma_0), \gamma_0)$. Moreover, $U^{(2)}_{\theta_P^2}(\mathbf{z}_P)$ is positive definite and $\theta_P(\gamma_0) = \theta_0$ for $P \supseteq P_0$.

Theorem 2 *Assume that **(C1), (C3-i)** and **(D)** hold. Then there are two positive constants δ_0^+ and η_0^+ such that if $c_n/n \leq \eta_0^+$, then*

$$M_n^+ \subseteq \left\{ \|T_n - \gamma_0\| \geq \delta_0^+ \sqrt{\frac{c_n}{n}} \right\} . \tag{11}$$

Therefore, higher smoothness of U yields a more explicit evaluation (11) of the overfitting set M_n^+. In particular, to overcome any overfitting, $\sqrt{c_n/n}$ must have at least the magnitude of $\|T_n - \gamma_0\|$. However, there is a price to pay for this new evaluation of M_n^+: it is in general less accurate than (10).

It is worth noting that the assumption **(C3-ii)** on the continuity of the minimization map $\gamma \mapsto \theta_P(\gamma)$ is not used in Theorem 2. Indeed, the smoothness **(D)** implies this continuity, as shown by Eq.(17) in the proof of Theorem 2 given below. As a consequence, under assumptions **(C1), (C2), (C3-i)** and **(D)**, both Theorems 1 and 2 apply.

2.2 Proofs.

Proof of Theorem 1. To describe M_n^- and M_n^+, we should first estimate

$$\Delta_n(P, P_0) = U_n(\widehat{\theta}_P) - U_n(\widehat{\theta}_0) = \sigma_P(T_n) - \sigma_{P_0}(T_n) = \xi_1 + \xi_2(P, P_0) + \xi_3 \tag{12}$$

where $\xi_1 = \sigma_P(T_n) - \sigma_P(\gamma_0)$, $\xi_2(P, P_0) = \sigma_P(\gamma_0) - \sigma_{P_0}(\gamma_0)$ and $\xi_3 = \sigma_{P_0}(\gamma_0) - \sigma_{P_0}(T_n)$, and then compare it with $n^{-1}c_n(p_0 - p)$ according to equations (3) and (4).

Evaluation of M_n^-. Let $P \not\supseteq P_0$. By (7), $\xi_2(P, P_0) \geq \Delta_2$. First assume $||T_n - \gamma_0|| < R_1$. Then by using the modulus of continuity Φ (5), we find

$$|\xi_1| \leq \Phi\left(||T_n - \gamma_0||\right), \qquad |\xi_3| \leq \Phi\left(||T_n - \gamma_0||\right) . \tag{13}$$

Assume in addition that $||T_n - \gamma_0|| < \Phi^{-1}\left(\frac{1}{2}\left[\Delta_2 - m\frac{c_n}{n}\right]\right)$. Then

$$\Phi\left(||T_n - \gamma_0||\right) < \frac{1}{2}\left[\Delta_2 - m\frac{c_n}{n}\right] .$$

Hence

$$\Delta_n(P, P_0) \geq \Delta_2 - 2\Phi\left(||T_n - \gamma_0||\right) > m\frac{c_n}{n} \geq (p_0 - p)\frac{c_n}{n}. \tag{14}$$

Consequently by (3), P should not be preferred to P_0, that is $\widehat{P}_n \notin \{P : P \not\supseteq P_0\}$. The evaluation (i) of M_n^- follows.

Evaluation of M_n^+. Let $P \supset P_0$ and $P \neq P_0$. Clearly, $\xi_2(P, P_0) = 0$. Again assume $||T_n - \gamma_0|| < R_1$. The estimates in (13) also hold with Ψ in place of Φ. Assume in addition that $||T_n - \gamma_0|| < \Psi^{-1}\left(\frac{1}{2}\frac{c_n}{n}\right)$. Then

$$\Psi\left(||T_n - \gamma_0||\right) < \frac{1}{2}\frac{c_n}{n} .$$

Hence

$$\Delta_n(P, P_0) \geq \ -2\Psi\left(||T_n - \gamma_0||\right) \ > -\frac{c_n}{n} \geq (p_0 - p)\frac{c_n}{n}. \tag{15}$$

Consequently, such a P should not be preferred to P_0. The evaluation (ii) of M_n^+ follows. ■

Proof of Theorem 2. Observe first that for each $P \subseteq M$, as $U^{(2)}_{\theta_P^2}(\theta, \gamma)$ is continuous and $U^{(2)}_{\theta_P^2}(\mathbf{z}_P)$ is positive definite, we may assume (by decreasing r_0 if necessary), that $U^{(2)}_{\theta_P^2}(\theta, \gamma)$ is positive definite everywhere on the ball $B(\mathbf{z}_P\ ; r_0)$. Let us define

$$b := \max_{P \subseteq M} \sup_{B_P} \left\{ \left\| \left[U^{(2)}_{\theta_P^2} \right]^{-1} \right\|, \ \left\| U^{(2)}_{\theta_P, \gamma} \right\| \right\} . \tag{16}$$

Step 1. Let us first prove that there is some $r_1 > 0$ such that we have for each $P \subseteq M$,

$$\text{If } ||\gamma - \gamma_0|| \leq r_1, \quad \text{then} \quad ||\theta_P(\gamma) - \theta_P(\gamma_0)|| \leq b^2||\gamma - \gamma_0||. \tag{17}$$

Fix some $P \subseteq M$ and assume that $||\gamma - \gamma_0|| < r_0$. Recall that $U^{(1)}_{\theta_P}(\theta_P(\gamma), \gamma) = U^{(1)}_{\theta_P}(\theta_P(\gamma_0), \gamma_0) = 0$. Since $U^{(2)}_{\theta_P^2}(\theta_P(\gamma_0), \gamma_0)$ is positive definite, an application of the implicit function theorem to $U^{(1)}_{\theta_P}$ says that there exists some ball $V_P = B(\gamma_0\,;\, r_{1,P})$ (we take $r_{1,P} \leq r_0$) and a continuously differentiable map $G : V_P \to \Theta_P$ such that if $\gamma \in V_P$, $U^{(1)}_{\theta_P}(\theta, \gamma) = 0$ if and only if $\theta = G(\gamma)$, with $(G(\gamma), \gamma) \in B(\mathbf{z}_P\,;\, r_0\,)$. In particular, if $||\gamma - \gamma_0|| \leq r_{1,P}$, then $\theta_P(\gamma) = G(\gamma)$. Since

$$G^{(1)}(\gamma) = -\left[U^{(2)}_{\theta_P^2}(G(\gamma), \gamma)\right]^{-1} U^{(2)}_{\theta_P, \gamma}(G(\gamma), \gamma)\,,$$

the estimate (17) follows from (16) taking $r_1 = \min\{r_{1,P} : P \subseteq M\}$.

Step 2. Let $P \supseteq P_0$. By definition, $\widehat{\theta}_0 := \hat{\theta}_{P_0}$,

$$\Delta_n(P, P_0) = U(\widehat{\theta}_P, T_n) - U(\widehat{\theta}_0, T_n) \geq U(\widehat{\theta}_P, T_n) - U(\theta_0, T_n).$$

Since $U^{(1)}_{\theta_P}(\widehat{\theta}_P) = 0$, applying Taylor's formula to the r.h.s. of the above inequality gives, for some $\widetilde{\theta} \in [\theta_0, \widehat{\theta}_P]$:

$$\Delta_n(P, P_0) \geq -\frac{1}{2}\,(\theta_0 - \widehat{\theta}_P)' U^{(2)}_{\theta_P^2}(\widetilde{\theta}, T_n)(\theta_0 - \widehat{\theta}_P).$$

Let us define

$$r_2 := r_1 \wedge (r_0/2) \wedge r_0/(2b^2), \quad \eta_0^+ = r_2^2\,, \quad \text{and} \quad \delta_0^+ = 1 \wedge \sqrt{2/b^5}\,. \tag{18}$$

Assume that $c_n/n \leq \eta_0^+$ and $||T_n - \gamma_0|| < \delta_0^+ \sqrt{\frac{c_n}{n}}$. By **(D),** $\theta_P(\gamma_0) = \theta_0$. It follows that $||T_n - \gamma_0|| \leq r_2 \leq \frac{1}{2} r_0$ and by (17), $||\widehat{\theta}_P - \theta_0|| \leq b^2 ||T_n - \gamma_0|| \leq \frac{1}{2} r_0$. Hence, the point $(\widehat{\theta}_P, T_n)$ as well as $(\widetilde{\theta}, T_n)$ belongs to B_P. So using (16), $||U^{(2)}_{\theta_P^2}(\widetilde{\theta}, T_n)|| \leq b$. Therefore, for such a choice of T_n and P, we obtain:

$$\Delta_n(P, P_0) \geq -\frac{1}{2} b ||\theta_0 - \widehat{\theta}_P||^2 \geq -\frac{1}{2} b^5 ||T_n - \gamma_0||^2 > -\frac{c_n}{n}. \tag{19}$$

Consequently, if $P \supset P_0$ and $P \neq P_0$, $\Delta_n(P, P_0) > -(p - p_0)\frac{c_n}{n}$. The new evaluation (11) of M_n^+ follows. ■

3 Applications

In this section, we shall illustrate Theorems 1 and 2 for several models.

3.1 Regression models with least squares estimation

Consider an univariate regression model

$$y_i = x_i'\theta + \varepsilon_i\,, \quad i = 1, \ldots, n, \;\; y_i \in \mathbb{R}, \qquad x_i \text{ and } \theta \in \mathbb{R}^m\,. \tag{20}$$

That is $Y = X\theta + \xi$, with $Y = (y_1, \cdots, y_n)'$, $X = (x_1, \cdots, x_n)'$ and $\xi = (\varepsilon_1, \cdots, \varepsilon_n)'$. Set $Q = \frac{1}{n} X'X$ and assume

$$(\mathbf{ML}) \;:\; Q \text{ is positive definite}\,. \tag{21}$$

For any positive definite matrix A, we set $||u||_A^2 := u'Au$ and denote its largest and smallest eigenvalues by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively. The least squares function is:

$$U_n(\theta) = n^{-1}||Y - X\theta||^2 = u_n + ||\theta - T_n||_Q^2 \tag{22}$$

with $u_n = n^{-1}\left[Y'Y - Y'X(X'X)^{-1}X'Y\right]$ and $T_n = (X'X)^{-1}X'Y$. Here $U(\theta,\gamma) = ||\theta - \gamma||_Q^2$, $F = \Theta = \mathbb{R}^m$ and $\gamma_0 = \theta_0$.

For $B \subseteq M$ and a matrix Γ (or vector) indexed by M, let Γ_B denote the restriction of Γ on B. Then, the submatrix $Q_B = \frac{1}{n} X_B'X_B$ is still positive definite.

Assumptions **(C1)** and **(C2)** are obviously satisfied. Furthermore, for each $P \subseteq M$, the (unique) minimum map is $\gamma \mapsto \theta_P(\gamma) = [\ \gamma_i 1_{i\in P}]$ defined for all $\gamma \in \mathbb{R}^m$. Thus **(C3)** holds with $R_1 = \infty$. It follows that $\sigma_P(\gamma) = ||\gamma_S||_{Q_S}^2$ where $S := M\backslash P$. Therefore $\sigma_P(\gamma_0) = \sigma_P(\theta_0) = ||\theta_{0,S}||_{Q_S}^2 = ||\theta_{0,P_0\backslash P}||_{Q_{P_0\backslash P}}^2$. Hence Δ_2 is equal to

$$\Delta_2 = \min_{P \not\supseteq P_0} ||\theta_{0,P_0\backslash P}||_{Q_{P_0\backslash P}}^2 .$$

We now estimate the modulus of continuity Φ and Ψ. For overmodels $P \supseteq P_0$, $\sigma_P(\gamma_0) = 0$. Since $\gamma_{0,S} = 0$,

$$\begin{aligned} \sigma_P(\gamma) - \sigma_P(\gamma_0) &= ||\gamma_S||_{Q_S}^2 = ||\gamma_S - \gamma_{0,S}||_{Q_S}^2 \\ &\le ||\gamma - \gamma_0||_Q^2 \le \lambda_{\max}(Q)\ ||\gamma - \gamma_0||^2 . \end{aligned}$$

Therefore we can take $\Psi(u) = \lambda_{\max}(Q)\, u^2$.

Next consider submodels $P \not\supseteq P_0$. Assume that $\theta_0 = \gamma_0 \neq 0$ and $||\gamma - \gamma_0|| \le u$.

$$\begin{aligned} |\sigma_P(\gamma) - \sigma_P(\gamma_0)| &= \left|\ ||\gamma_S||_{Q_S}^2 - ||\gamma_{0,S}||_{Q_S}^2\ \right| \le ||\gamma_S + \gamma_{0,S}||_{Q_S}||\gamma_S - \gamma_{0,S}||_{Q_S} \\ &\le \lambda_{\max}(Q_S)\, ||\gamma_S + \gamma_{0,S}||\, ||\gamma_S - \gamma_{0,S}|| \le \lambda_{\max}(Q)\ ||\gamma + \gamma_0||\, ||\gamma - \gamma_0|| \\ &\le \lambda_{\max}(Q)\, u(u + 2||\gamma_0||) . \end{aligned}$$

Thus we can take $\Phi(u) = \lambda_{\max}(Q)\, u(u + 2||\gamma_0||)$ for submodels. A straightforward application of Theorem 1 yields

Proposition 3 *For the regression (20) and the least squares function (22), assume that* ***(ML)*** *holds. Then*

(i) Let $f_0 := \sqrt{\frac{1}{4}\Delta_2/\lambda_{\max}(Q) + ||\theta_0||^2} - ||\theta_0||$. *If* $\theta_0 \neq 0$ *and* $c_n/n \le \Delta_2/(2m)$, *have*

$$M_n^- \subseteq \left\{ \left\|(X'X)^{-1}X'Y - \theta_0\right\| \ge f_0 \right\} . \tag{23}$$

(ii) For the overfitting set, we have

$$M_n^+ \subseteq \left\{ \left\|(X'X)^{-1}X'Y - \theta_0\right\| \ge \sqrt{\frac{1}{2\lambda_{\max}(Q)}\frac{c_n}{n}} \right\} . \tag{24}$$

It is worth noting that the matrix Q, hence Δ_2 and $\lambda_{\max}(Q)$, depend on the sample size n. For Δ_2, note that

$$\Delta_2 \ge \lambda_{\min}(Q) \min_{i\in P_0} \theta_{0,i}^2 .$$

Therefore any asymptotic analysis will depend on the behavior of both $\lambda_{\min}(Q)$ and $\lambda_{\max}(Q)$.

Remark. The estimation of the support of the mean of a multidimensional variable can be treated in a similar way. Let $Y_1, \cdots, Y_n$ be n i.i.d. m-dimensional observations with mean $\theta_0 = \mathbb{E}Y_j$. The aim is to estimate the *support* $P_0 = \{i : \theta_{0,i} \neq 0\}$ of θ_0. Consider the least squares function

$$U_n(\theta) = n^{-1} \sum_{j=1,n} ||Y_j - \theta||^2 = ||\theta - \bar{Y}||^2 + n^{-1} \sum_{j=1,n} ||Y_j - \bar{Y}||^2, \tag{25}$$

with $\overline{Y} = n^{-1}(Y_1 + \cdots + Y_n)$. Analogously to the regression case, we have the following

Proposition 4

(i) *Let* $e_0 := \sqrt{\frac{1}{4}\Delta_2 + ||\theta_0||^2} - ||\theta_0||$. *If* $\theta_0 \neq 0$ *and* $c_n/n \leq \Delta_2/(2m)$ *, we have*

$$M_n^- \subseteq \left\{ ||\overline{Y} - \theta_0|| \geq e_0 \right\} . \tag{26}$$

(ii) *For the overfitting set, we have*

$$M_n^+ \subseteq \left\{ ||\overline{Y} - \theta_0|| \geq \sqrt{\frac{1}{2}\frac{c_n}{n}} \right\} \tag{27}$$

3.2 AR on $\mathbb{Z}$ or Markov field on $\mathbb{Z}^d$ with Whittle's pseudo-likelihood

AR(m) process. For a positive integer m and $M = \{1, 2, \cdots, m\}$, define

$$\Theta = \{\theta \in \mathbb{R}^m : 1 - \sum_{\ell=1}^{m} \theta_\ell \, z^\ell \neq 0 \quad \text{for } z \in \mathbb{C},\ |z| \leq 1 \} . \tag{28}$$

For $\theta \in \Theta$ and some white noise ε, we consider a causal univariate AR(m) process

$$X_t = \sum_{\ell=1}^{m} \theta_\ell X_{t-\ell} + \varepsilon_t,\ t \in \mathbb{Z} \tag{29}$$

Conditional Markov field [CAR(M)] model ([25], [10]). Let M be some finite subset of $(\mathbb{Z}^d)^+$, the positive half space of $\mathbb{Z}^d$ with respect to the lexicographic order. Denote by m the size of M and define

$$\Theta = \{\theta \in \mathbb{R}^m \ : \ 1 - 2 \sum_{\ell \in M} \theta_\ell \cos\langle \lambda, \ell \rangle \ > 0 \ , \quad \text{for all } \lambda \in \mathbb{T}^d\} \tag{30}$$

with $\mathbb{T} = [0, 2\pi[$. If $\theta \in \Theta$, the following equations

$$X_t = \sum_{\ell \in M} \theta_\ell (X_{t+\ell} + X_{t-\ell}) + e_t, \quad \text{with } E(e_t X_u) = 0 \ \text{ for } t \neq u \tag{31}$$

defines a CAR(M) Markov field with support M.

Spectral densities and Whittle's pseudo-likelihood

The spectral density f_θ of both models (29) and (31) takes the form

$$f_\theta^{-1}(\lambda) = (2\pi)^d \kappa^{-1}\{c_0(\theta) + 2\sum_M c_\ell(\theta)\cos\langle\lambda,\ell\rangle\}\ . \tag{32}$$

For the AR(m), $\kappa = \sigma_\varepsilon^2$, $c_\ell(\theta) = \theta_0\theta_\ell + \cdots + \theta_{m-\ell}\theta_m$ (in this formula, $\theta_0 = -1$) for $\ell = 0, \ldots, m$; and for the CAR(M) model, $\kappa = \sigma_e^2$, $c_0(\theta) = 1$, $c_\ell(\theta) = -\theta_\ell$ for $\ell \in M$.

Suppose that the process is observed on a rectangle $[\mathbf{1}, \mathbf{n}] := [1, n_1] \times \cdots \times [1, n_d]$ of $\mathbb{Z}^d$ ($d = 1$ for the AR(m) process). The Gaussian pseudo-likelihood (Whittle, [30]) is given by

$$\begin{cases} U_n(\theta) = \log\sigma_\varepsilon^2 + \sigma_\varepsilon^{-2}\{c_0(\theta)\widehat{\gamma}_n(0) + 2\sum_{s=1}^m c_s(\theta)\widehat{\gamma}_n(s)\}, \quad \text{for AR}(m) \\ U_n(\theta) = (2\pi)^{-d}\int_{\mathbb{T}^d}\log f_\theta(\lambda)d\lambda + \sigma_e^{-2}\{\widehat{\gamma}_n(0) - 2\sum_{s=\in M}\theta_s\widehat{\gamma}_n(s)\}, \quad \text{for CAR}(M) \end{cases} \tag{33}$$

where $\widehat{\gamma}_n(s)$ denotes the sample covariance $n^{-1}\sum_{t\in[\mathbf{1},\mathbf{n}]} X_t X_{t+s}$ with the convention $X_u = 0$ if $u \notin [\mathbf{1}, \mathbf{n}]$. By taking $T_n = (\widehat{\gamma}_n(u), u \in M \cup \{0\})$, we can factorise U_n with the function

$$\begin{cases} U(\theta,\gamma) = \log\sigma_\epsilon^2 + \sigma_\epsilon^{-2}\{c_0(\theta)\gamma(0) + 2\sum_{s=1}^m c_s(\theta)\gamma(s)\}, \quad \text{for AR}(m) \\ U(\theta,\gamma) = (2\pi)^{-d}\int_{\mathbb{T}^d}\log f_\theta(\lambda)d\lambda + \sigma_e^{-2}\{\gamma(0) - 2\sum_{s=\in M}\theta_s\gamma(s)\}, \quad \text{for CAR}(M). \end{cases} \tag{34}$$

Here the state space F is the collections of all $(m+1)$ Fourier coefficients $\gamma = [\widehat{h}(\ell)]$ with $\ell \in M \cup \{0\}$, h running through the set of positive and Lebesgue integrable functions on $\mathbb{T}^d$. We take $\gamma_0 = \left[\widehat{(2\pi)^d f_{\theta_0}}(\ell)\right]$, the Fourier coefficients of $(2\pi)^d f_{\theta_0}$ on $M \cup \{0\}$.

Proposition 5 *For both the AR(m) process defined in (29)-(28) and the CAR(M) field defined in (31)-(30) with the Whittle pseudo-likelihood (33), Conditions* ***(C)*** *and* ***(D)*** *are fulfilled and Theorems 1 and 2 apply.*

The proof is given in Appendix A.

3.3 Likelihood for categorical data, conditional pseudo-likelihood for Markov field

Let X be a random variable with a finite state space $E = \{a_0, a_1, a_2, \cdots, a_K\}$, $K \geq 1$, the distribution of this variable being conditional to some $v \in V = \{v_1, v_2, \cdots, v_L\}$. In econometric models, v is some conditioning exogenous variable, while for a Markov field $X = (X_i)$, $v = X_{\partial i}$ represents a neighborhood configuration around some site i. Natural estimating functions include the likelihood and Besag's conditional pseudo-likelihood, see [5, 10]. Assume that such a conditional distribution is defined by

$$\mathbb{P}(X = a|v) = \pi_\theta(a\,|v) = \frac{\exp\langle\theta, \boldsymbol{\phi}(a,v)\rangle}{\sum_{x\in E}\exp\langle\theta,\boldsymbol{\phi}(x,v)\rangle} \tag{35}$$

with $\theta \in \Theta = \mathbb{R}^m$. Here $\boldsymbol{\phi}(a, v) = [\phi_\ell(a, v)]$, $\ell = 1, \ldots, m$, are *conditional potentials*. To ensure the identifiability of the model, we take $\boldsymbol{\phi}(a_0, v) = 0$.

Suppose that for each v we have $n_v \geq 1$ independent observations $\{X_{iv} = (X_i|v)\}_{1 \leq i \leq n_v}$ under v. The conditional log-likelihood of the observations is

$$\mathcal{L}_n(\theta) = \sum_v \sum_{i=1,n_v} \log \pi_\theta(X_{iv}|v) = \sum_{a \in E, v \in V} n_{av} \log \pi_\theta(a|v) \tag{36}$$

where $n_{av} = \#\{i : X_{iv} = a\}$, $n_v = \sum_a n_{av}$ and $n = \sum_v n_v$. Thus we have for $U_n = -n^{-1}\mathcal{L}_n$

$$\begin{cases} U_n(\theta) = U(\theta, T_n) = -n^{-1} \sum_v n_v \sum_a T_n(a, v) \log \pi_\theta(a|v) \\ \text{with} \\ \quad T_n = (T_n(a, v) : a \in E, v \in V), \quad T_n(a, v) = n_{av}/n_v \ . \end{cases} \tag{37}$$

Here $\Theta = \mathbb{R}^m$, and F is the set of all conditional distributions induced by (35): an element $\gamma \in F$ is a collection of L conditional distributions $\gamma = (\pi_\theta(a|v))$, $(a, v) \in E \times V$, for some $\theta \in \mathbb{R}^m$. Also we take $\gamma_0 = (\pi_{\theta_0}(a, v))$. Let us define the $m \times (K+1)L$ matrix $\Sigma = (\phi_\ell(a, v))$, for row index $\ell = 1, \ldots, m$ and column index $(a, v) \in E \times V$.

Proposition 6 *Consider the model (35) with the pseudo-likelihood function (37). Under the condition*

$$\textbf{\textit{(CAT)}} : \qquad \textit{the matrix } \Sigma \textit{ is of full rank } m \ , \tag{38}$$

Conditions ***(C)*** *and* ***(D)*** *are both fulfilled and Theorems 1 and 2 apply.*

The proof is given in Appendix B.

Let us give two examples where **(CAT)** is satisfied.

Example 1: logistic regression. Here we assume $v \in R^q$ for some integer $q > 0$. The polytomic logistic regression model takes the form :

$$\mathbb{P}(X = a_i|v) = \frac{\exp\langle \beta_i, v \rangle}{\sum_{s=0}^K \exp\langle \beta_s, v \rangle} \ , \quad i = 0, 1, \ldots, K$$

with $\beta_0 = 0$. The parameters are $\theta = (\beta_1', \ldots, \beta_K')' \in \mathbb{R}^{qK}$ and $\boldsymbol{\phi}(a_i, v) = (0', \ldots, 0', v', 0', \ldots, 0')'$ where v is at position i. Assume that V spans $\mathbb{R}^q$. Then Σ spans R^{qK} and **(CAT)** is satisfied (with $m = qK$). □

Example 2: Markov field on $\mathbb{Z}^d$ [10]. Let $M = \{u_1, u_2, \cdots, u_m\}$ be m sites of the positive half space $(\mathbb{Z}^d)^+$ with $u_1 = 0$. The set M defines a neighborhood relation: $i \sim j$ if $i - j$ or $j - i$ belongs to $M \backslash \{0\}$. For simplicity, consider a homogeneous field with a singleton potential $\Psi_1 : E \to R$, and pair potentials $\Psi_\ell : E \times E \to R$, defined for any pair of sites (i, j) satisfying $i - j = \pm u_\ell$, $\ell = 2, \ldots, m$. Thus the conditional distribution at site i is defined by the conditional energy $\langle \theta, \boldsymbol{\phi}(x_i, x_{\partial i}) \rangle$, where $\theta = (\theta_1, \theta_2, \cdots, \theta_m)' \in \mathbb{R}^m$, $\boldsymbol{\phi} = (\phi_\ell, \ \ell = 1, m)'$, $\partial i = \{i \pm u_\ell, \ell = 2, \ldots, m\}$ and $\phi_1(x_i, x_{\partial i}) = \Psi_1(x_i)$, $\phi_\ell(x_i, x_{\partial i}) = \Psi_\ell(x_i, x_{i+u_\ell}) + \Psi_\ell(x_i, x_{i-u_\ell}), \ell = 2, m$.

To specify, consider an Ising model: $E = \{-1, 1\}$, $\Psi_1(x) = x$, $\Psi_\ell(x, y) = xy$ for $\ell = 2, \ldots, m$. Define $v_\ell = x_{i+u_\ell} + x_{i-u_\ell}$ for $\ell = 2, \ldots, m$, and $\Delta(v) = {}^t(1, v_2, v_3, \cdots, v_m)$. It is easy to see that $\boldsymbol{\phi}(x_i, x_{\partial i}) = x_i \Delta(v)$. Hence if $\{\Delta(v), v \in V\}$ spans $\mathbb{R}^m$, the condition **(CAT)** is satisfied. □

4 Upper bounds for the misfitting probabilities $\mathbb{P}(M_n^-)$, $\mathbb{P}(M_n^+)$

We now give upper bounds for $\mathbb{P}(M_n^-)$ and $\mathbb{P}(M_n^+)$ for the examples considered in the previous section and two infinite variance models. Even we do not state it explicitly, sufficient conditions for the weak consistency can be straightforwardly derived from these upper bounds.

Probability and expectation under θ_0 will be denoted by $\mathbb{P}_0$ and $\mathbb{E}_0$ respectively. The main goal is to evaluate deviation probabilities like $\mathbb{P}_0\{||T_n - \gamma_0|| \geq a_n\}$, a_n being a constant in the case of M_n^-, while for M_n^+, a_n is proportional to $\sqrt{c_n/n}$ which usually tends to zero as $n \to \infty$. Consequently, $\mathbb{P}(M_n^-)$ is related to large deviations of T_n and $\mathbb{P}(M_n^+)$ to its moderate deviations.

We follow an approach based on exponential inequalities. The main interest of this approach is that large or moderate deviation probabilities can be treated in an unified way. However, a possible drawback of this approach could be that the constants involved in the upper bounds given below may be not optimal.

A common fact is that $\mathbb{P}(M_n^-)$ vanishes exponentially fast (provided c_n/n is small), while $\mathbb{P}(M_n^+)$ decreases at a slower rate depending on the magnitude of c_n/n. For instance, for the BIC rate $c_n = \log n$, $\mathbb{P}(M_n^+)$ is of polynomial order $O(n^{-\alpha})$ for some $\alpha > 0$.

Let us recall the exponential inequalities used. We shall say that a zero-mean, real-valued variable X has *an exponential moment* $\mathbf{E}(\tau, g)$ if the following condition is fulfilled for some positive constants τ and g,

$$\mathbf{E}(\tau, g) \ : \ \mathbb{E}\, e^{tX} \leq e^{\frac{1}{2}gt^2} \ \text{ for } \ |t| \leq \tau \ . \tag{39}$$

This is equivalent to the following moment condition (see Lemma 2.2 in [21]) :

$$\exists\, a > 0, \quad \mathbb{E}e^{a|X|} < \infty \ . \tag{40}$$

Moreover, any family $\mathcal{F} = \{X_\alpha\}$ of real variables will be said to have an *uniform exponential moment* $\mathbf{E}(\tau, g)$ if (39) is satisfied for each $X_\alpha \in \mathcal{F}$ with some uniform constants τ and g.

Furthermore for any sequence of independent and centered real variables $(X_n)_{n\geq 1}$ having a uniform exponential moment $\mathbf{E}(\tau, g)$, the following exponential bound holds for the mean $\overline{X}_n = n^{-1}\sum_{j=1}^n X_j$ (see Theorem 2.6 in [21])

$$\mathbb{P}\left(|\overline{X}_n| \geq x\right) \leq 2\exp\left\{-n\left(\frac{x^2}{2g} \wedge \frac{\tau x}{2}\right)\right\} \ . \tag{41}$$

4.1 Regression models

For the regression model (20) defined in Section 3.1, we will assume that

$$\begin{cases} \text{The variables } (\varepsilon_j)_{j\geq 1} \text{ are zero-mean, independent and} \\ \\ \text{have an uniform exponential moment } \mathbf{E}(\tau^*, g^*) \text{ for some } \tau^*, g^* > 0. \end{cases} \tag{42}$$

As the sample size n may vary, we add the subscript n to previously defined variables X, Y, ξ and Q. Assume also the following condition on the exogenous variables $\{x_i\}$

$$A^* := \sup\left\{x_i^2(\ell) : i \geq 0,\, 1 \leq \ell \leq m\right\} \ < \ \infty \ , \tag{43}$$

$$\lambda_* := \inf_{n\geq m} \lambda_{\min}(Q_n) \ > \ 0 \, . \tag{44}$$

It is worth noting that for an i.i.d. process (ε_i), (43) can be weakened, see [31] and [22].

Proposition 7 *In the framework of Section 3.1, assume that (42), (43) and (44) are satisfied. Then there exist positive constants Δ_*, D_1, D_2 and D_3 such that*

(i) If $\theta_0 \neq 0$ and $c_n/n \leq \Delta_*/(2m)$, we have

$$\mathbb{P}_0(M_n^-) \leq 2me^{-D_1 n} \ .$$

(ii) For the overfitting set, we have

$$\mathbb{P}_0(M_n^+) \leq 2m \exp\left\{-\left[D_2\, c_n \wedge D_3\sqrt{n c_n}\right]\right\} \ .$$

Proof. Easy calculus give for the constants involved in Proposition 3

$$\begin{aligned} \lambda_{\max}(Q_n) &\leq mA^* \ , & (45)\\ \Delta_2 &\geq \Delta_* := \lambda_* \min_{i\in P_0} \theta_{0,i}^2 \ , & (46)\\ f_0 &\geq f_* := \sqrt{\frac{\Delta_*}{4mA^*} + ||\theta_0||^2} - ||\theta_0|| \ . & (47) \end{aligned}$$

As Q_n is always positive definite for $n \geq m$, Condition **(ML)** is satisfied. Let us use the evaluations (23)-(24). Since $Y_n = X_n\theta_0 + \xi_n$, we get $(X_n'X_n)^{-1}X_n'Y_n - \theta_0 = n^{-1}Q_n^{-1}X_n'\xi_n$. Let Z_n be the m-dimensional vector $n^{-1}\ X_n'\xi_n$. Thus for $u \geq 0$,

$$\left\{\left\|(X_n'X_n)^{-1}X_n'Y_n - \theta_0\right\| \geq u\right\} \subset \left\{\|Z_n\| \geq u\sqrt{\lambda_{\min}(Q_n)}\right\} \subset \left\{\|Z_n\| \geq u\sqrt{\lambda_*}\right\}.$$

Furthermore, the ℓ-th component of Z_n is $Z_n(\ell) = n^{-1}\sum_{i=1}^n x_i(\ell)\varepsilon_i$. For any t such that $|t| \leq \tau^*/\sqrt{A^*}$, since $|tx_i(\ell)| \leq \tau^*$, we find by (42) and (43),

$$\mathbb{E}e^{tx_i(\ell)\varepsilon_i} \leq \exp\left[\frac{1}{2}g^*x_i^2(\ell)t^2\right] \leq \exp\left[\frac{1}{2}g^*A^*t^2\right] \ .$$

Hence the weighted sequence $\{x_i(\ell)\varepsilon_i\}_{i\geq 1}$ has a uniform exponential moment $\mathbf{E}(\tau^*/\sqrt{A^*}, g^*A^*)$. Since $\left\{||Z_n|| \geq u\sqrt{\lambda_*}\right\} \subset \bigcup_{\ell=1}^m \left\{|Z_n(\ell)| \geq u\sqrt{\lambda_*}/\sqrt{m}\right\}$, we have

$$\mathbb{P}\left(\|Z_n\| \geq u\sqrt{\lambda_*}\right) \leq 2m\exp\left\{-n\left(\frac{\lambda_* u^2}{2g^*mA^*} \wedge \frac{\tau^* u}{2}\sqrt{\frac{\lambda_*}{mA^*}}\right)\right\} \ .$$

The results follows by taking $u = f_*$ and $u = \left\{(2mA^*)^{-1}c_n/n\right\}^{1/2}$ for $\mathbb{P}_0(M_n^-)$ and $\mathbb{P}_0(M_n^+)$, respectively. The involved constants are

$$D_1 = \frac{\lambda^* f_*^2}{2g^*mA^*} \wedge \frac{\tau^* f_*\sqrt{\lambda_*}}{2\sqrt{mA^*}} \ , \quad D_2 = \frac{\lambda_*}{2g^*(2mA^*)^2} \, , \quad D_3 = \frac{\tau^*\sqrt{\lambda_*}}{4\sqrt{2}mA^*} \, . \quad \blacksquare$$

4.2 AR and CAR Markov fields

For the AR(m) process (29), we shall assume the following

$$(\varepsilon_t) \text{ is a centered i.i.d. sequence satisfying for some } a > 0,\ \mathbb{E}e^{a\varepsilon_1^2} < \infty\ . \tag{48}$$

and for the CAR(M) field (31)

$$X = (X_t) \text{ is a centered Gaussian field.} \tag{49}$$

Both the AR(m) process and the Gaussian CAR(M) field are linear processes, i.e.

$$X_t = \sum_{s \geq 0} a_s \varepsilon_{t-s} \tag{50}$$

where (ε_t) is the corresponding innovation process (for the CAR(M) case, see e.g. Theorem 1.2.3 in [10]). Moreover, there is some $\alpha \in [0,1)$ and $A \geq 0$ such that for $s = (s_1, \ldots, s_d) \in \mathbb{Z}^d$, $|a_s| \leq A\alpha^{|s_1|+\cdots+|s_d|}$. In the AR($m$) case, $d = 1$ and (ε_t) is the i.i.d. sequence defined in (29), while in the Gaussian CAR(M) case, ε_t is given by $\varepsilon_t = \mathbb{E}\,(X_t | X_s,\ s < t)$ with $<$ the lexicographic order on $\mathbb{Z}^d$. In the CAR(M) case, the variables (ε_t) are Gaussian and uncorrelated, hence independent, and the moment condition $\mathbb{E}e^{a\varepsilon_1^2} < \infty$ holds for some $a > 0$.

The common linear representation (50) makes a unified analysis of both models possible. Recall that the CAR(M) model is observed on a rectangle $[\mathbf{1}, \mathbf{n}] = [1, n_1] \times \cdots \times [1, n_d]$ of size $n = n_1 \cdots n_d$.

Proposition 8 (1). The CAR(M) case: *Assume that (30) and (49) are satisfied. Then there are positive constants n_*, μ, μ' and σ_j, ν_j, ν_j' for $j = 1, 2$, such that for $n_i \geq n_*$, $i = 1, \ldots, d$,*

(i) If $c_n/n \leq \eta_0^-$, we have

$$\mathbb{P}(M_n^-) \leq \mu e^{-n\mu'}\quad .$$

(ii) If $nc_n \geq \sigma_1$ and $c_n/n \leq \sigma_2$, we have

$$\mathbb{P}(M_n^+) \leq \nu_1 \left\{ 1 + \log\frac{n}{c_n} + \left(\log\frac{n}{c_n}\right)^2 \right\} e^{-\nu_1' c_n} + \nu_2 \left\{ 1 + \sqrt{\frac{n}{c_n}} + \frac{1}{c_n} \right\} e^{-\nu_2'\sqrt{nc_n}}\ . \tag{51}$$

(2). The AR(m) case: *Assume that (28) and (48) are satisfied. Then the same conclusions hold.*

The proof is given in Appendix C where the constants are made explicit. These results show that once $c_n/n \leq \eta_0^-$, $\mathbb{P}(M_n^-)$ vanishes exponentially fast. Moreover, the upper bound for $\mathbb{P}(M_n^+)$ is close to optimal in the following sense. Consider the AR(m) case with $c_n = 2C \log\log n$ for some $C \geq 0$. The r.h.s of (51) is equivalent to $\nu_1 (\log n)^{2(1-\nu_1' C)}$. Therefore, $\mathbb{P}(M_n^+) \to 0$ if $C > 1/\nu_1'$, while if $C \leq 1/\nu_1'$, this upper bound does not goes to zero. The existence of such a critical value for C is already known for strong consistency: Theorem 5.4.1 of [15] shows that strong consistency holds if and *only if* $C > 1$. If we conjectured the same critical value for weak consistency, the Hannan-Deistler's result seems to indicate that the upper bound in (51) is nearly optimal although we do not recover the exact critical value $C = 1$.

4.3 Categorical data models and finite state space Markov fields

Within the framework of Section 3.3, probability estimates for M_n^- and M_n^+ will be based on the following result (see [16, 21]): Given n independent real variables $(X_k)_{1\le k\le n}$, each of them having a compact range $[a_k, b_k]$ ($a_k < b_k$), the following deviation estimate holds for the average $\overline{X}_n$

$$\mathbb{P}(|\overline{X}_n| \ge x) \le 2\exp -\frac{2n^2x^2}{\sum_{k=1}^n (b_k - a_k)^2} \ . \tag{52}$$

On the other hand, since

$$\{||T_n - \gamma_0|| \ge x\} \subseteq \bigcup_{(a,v)\in E\times V} \left\{|T_n(a,v) - \pi_{\theta_0}(a|v)| \ge \frac{x}{KL}\right\},$$

and $T_n(a,v) = n_v^{-1} n_{av} \in [0,1]$, we have by (52)

$$\mathbb{P}(||T_n - \gamma_0|| \ge x) \le 2K \sum_V \exp - \left\{2n_v \left(\frac{x}{KL}\right)^2\right\} \ .$$

For the constants δ_0^-, η_0^- and δ_0^+ defined in Theorem 1 and 2, set

$$c^- = 2[\delta_0^-/(KL)]^2, \quad c^+ = 2[\delta_0^+/(KL)]^2 \quad .$$

A straightforward application of Proposition 6 yields

Proposition 9 *Under the condition (38), we have*

(i) If $c_n/n \le \eta_0^-$, then

$$\mathbb{P}(M_n^-) \le 2K \sum_{v\in V} e^{-c^- n_v}.$$

(ii) Without any condition on c_n,

$$\mathbb{P}(M_n^+) \le 2K \sum_{v\in V} \exp\left[-c^+ \left(\frac{n_v}{n}\right) c_n\right] \ .$$

On the (CAT) condition for Markov field : The condition**(CAT)** (38) requires that $n_v \ge 1$ for each v in some subset V of the neighborhood configurations. Here $(n_v)_{v\in V}$ are random. However, if X is observed on $[\mathbf{1}, \mathbf{n}] = [1,n]^d$, this requirement will be fulfilled almost surely for large n, thanks to the following subergodicity result [8]:

$$\exists \alpha > 0, \text{ such that almost surely, } \forall v \in V, \ \liminf_{n\to\infty} \frac{n_v}{n^d} \ge \alpha \ .$$

4.4 Models with infinite variances

Once c_n/n is small enough and following Theorem 1 and 2, the misfitting set $M_n = M_n^- \cup M_n^+$ can be estimated as $M_n \subseteq \left\{||T_n - \gamma_0|| > \delta\sqrt{c_n/n}\right\}$ for some $\delta > 0$. We shall estimate $\mathbb{P}_0(M_n)$ for two models involving variables with infinite variance.

4.4.1 Sample from a α-stable law with exponent $\alpha \in (1,2)$

Let us consider an i.i.d. sample of an m-dimensional random vector $Y = \theta_0 + \varepsilon$. Assume that each component $\varepsilon(j)$ of ε is a symmetric α-stable variable with index $\alpha \in (1,2)$ (see [21], Chap. 3 for more reference on stable variables). Such a α-stable variable Z satisfies : (1). as $x \to \infty$, $x^\alpha \mathbb{P}(|Z| > x) \to C$, where C is a characteristic constant; (2). For any sample $(Z_i)_{i=1,\dots,n}$, the normalized sample mean $n^{-1/\alpha} \sum_{i=1}^n Z_i$ has the same α-stable distribution as Z. Note that since $\alpha > 1$, $(2-\alpha)/\alpha < 1$. Thus straightforward application of (9) and (11) yields the following result.

Proposition 10 *Assume that*

$$\frac{c_n}{n} \to 0 \quad \textit{and} \quad \frac{c_n}{n^{(2-\alpha)/\alpha}} \to \infty \; . \tag{53}$$

Then there is a positive constant D such that for large enough n

$$\mathbb{P}_0(M_n) \leq D \left[\frac{c_n}{n^{(2-\alpha)/\alpha}} \right]^{-\alpha/2} \; . \tag{54}$$

4.4.2 Infinite variance AR(m) process

Consider an AR(m) process (X_t) as defined in (29)-(28), where (ε_t) are i.i.d. with a common distribution in the domain of attraction of a symmetric stable distribution with index $\alpha \in (0,2)$. For such a process, expectations of sample auto-covariances are undefined. Therefore the Whittle PL (33) is no longer useful.

However, expectations of the sample autocorrelations $\widehat{\rho}_n(s) := \widehat{\gamma}_n(s)/\widehat{\gamma}_n(0)$ are well-defined [14], and converge to

$$\rho_0(s) = \frac{\sum_{j \geq 0} b_j b_{j+s}}{\sum_{j \geq 0} b_j^2} \; , \tag{55}$$

where (b_j) are the coefficients of the linear representation of the process: $X_t = \sum_{s \geq 0} b_s \varepsilon_{t-s}$.

For estimation purpose, we modify the Whittle PL (33) as follows (still denoted U_n)

$$U_n(\theta) = c_0(\theta)\widehat{\rho}_n(0) + 2\sum_{s=1}^m c_s(\theta)\widehat{\rho}_n(s) = \left(\sum_{t=1}^n X_t^2 \right)^{-1} \cdot \sum_{t=1}^n \left(X_t - \sum_{s=1}^m \theta_s X_{t-s} \right)^2 \; ,$$

with the same $c_s(\theta)$ as in (33), $T_n = [\widehat{\rho}_n(s)]$ and $\gamma_0 = [\rho_0(s)]$ where $1 \leq s \leq m$. Assumptions **(C1), (C2), (C3-i)** and **(D)** still hold (as in the finite variance case, Proposition 5). Hence Theorem 1 and 2 apply.

To control the wrong fitting probabilities, let us recall the following Central Limit Theorem on T_n, proved in [7] (see also [19])

$$\left(\frac{n}{\log n} \right)^{1/\alpha} [\rho_n(s) - \rho_0(s)] \longrightarrow \mathcal{L}(s), \qquad \mathbb{P}_0\text{-weakly}, \tag{56}$$

where $\mathcal{L}(s)$ is some limiting distribution. Consequently, an application of (9) and (11) yields

Proposition 11 *Assume that*

$$\frac{c_n}{n} \to 0 \quad \textit{and} \quad \left(\frac{n}{\log n}\right)^{1/\alpha} \left(\frac{c_n}{n}\right)^{1/2} \longrightarrow \infty \ . \tag{57}$$

Then

$$\mathbb{P}_0(M_n) \to 0 \ . \tag{58}$$

For instance, the Akaike's information criterion ($c_n \equiv 2$) is consistent: we have recovered a result proved by [6] and [18].

5 Strong consistency of the model selection criterion

This section is devoted to illustrate how strong consistency can be derived from Theorems 1 and 2. To this end, assume that the following upper bound is available on the a.s. convergence rate of the statistic T_n

$$\exists A > 0, \ \limsup \left(\frac{n}{2\log\log n}\right)^{1/2} ||T_n - \gamma_0|| \leq A \ \text{ a.s.} \tag{59}$$

In such a case, strong consistency holds.

Theorem 12 *Assume that (59) and the evaluations (9) and (11) hold. Then*

(i) *If* $\limsup c_n/n < \eta_0^-$, *almost surely, for large enough* n, *underfitting is impossible, that is* $M_n^- = \emptyset$.

(ii) *If* $\limsup c_n/n < \eta_0^+$, *and* $\liminf \ c_n/(2\log\log n) > (A/\delta_0^+)^2$, *then almost surely, for large enough* n, *overfitting is impossible, that is* $M_n^+ = \emptyset$.

Proof. First note that (59) ensures $||T_n - \gamma_0|| \to 0$. Hence a.s. the set $\{||T_n - \gamma_0|| \geq \delta_0^-\}$ is empty for large enough n, and so is M_n^- by (9).

For M_n^+, we know, by (ii) of Theorem 2, that $M_n^+ \subseteq \{||T_n - \gamma_0|| \geq \delta_0^+ \sqrt{c_n/n}\} =: W_n$. On $W := \limsup W_n = \{||T_n - \gamma_0|| \geq \delta_0^+ \sqrt{c_n/n}$ infinitely often$\}$, there is some subsequence $n_k \uparrow \infty$, such that $||T_n - \gamma_0|| \geq \delta_0^+ \sqrt{c_{n_k}/n_k}$. Hence on W

$$\limsup \left(\frac{n}{2\log\log n}\right)^{1/2} ||T_n - \gamma_0|| \geq \delta_0^+ \liminf \left(\frac{c_n}{2\log\log n}\right)^{1/2} > A.$$

By (59), W is negligible. The result (ii) follows. ■

It should be pointed out that (59) is a natural assumption when considering the strong consisitency. So, Theorem 12 just recovers this well-known fact. Again, if we remember the Hannan-Deistler's result (Theorem 5.4.1 of [15]) applied to AR models (see also comments at the end of Section 4.2), our conditions on the penelization rate c_n in Theorem 12 are optimal up to some constant factor in that case. We have lost some precision, but Theorem 12 can be applied to many other models than the AR ones.

ACKNOWLEGDEMENT. *The authors are grateful to the anonymous referees for their helpful comments which have contributed to improve the results of the paper.*

References

[1] H. Akaike. Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.*, 21:243–247, 1969.

[2] Z. D. Bai, K. Subramanyan, and L. C. Zhao. On determination of the order of an Autoregressive Model. *J. Multivariate Analysis*, 27:40–52, 1988.

[3] Z.D. Bai, P.R. Krishnaiah, N. Sambamoorthi, and L.C. Zhao. Model selection for non-linear models. *Sankhya: Indian J. Statist., Ser. B*, 54:200–219, 1992.

[4] Z.D. Bai, P.R. Krishnaiah, and L.C. Zhao. On rates of convergence of efficient detection criteria in signal processing with white noise. *IEEE Trans. Inform. Theory*, 35(2):380–388, 1989.

[5] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. B*, 36:192–236, 1974.

[6] R. J. Bhansali. Consistent order determination for processes with infinite variance. *J. Roy. Statist. Soc. B*, 50:46–60, 1988.

[7] R. Davis and S. Resnick. Limit theory for the sample covariance and correlation functions of moving averages. *Ann. Statist.*, 14:533–558, 1986.

[8] S. Geman and C. Graffigne. Markov random fields image models and their applications to computer vision. In M. Gleason, editor, *Proc. Int. Congress Math.* A.M.S. Providence, 1986.

[9] J. Geweke and R. Richard. Estimating regression models of finite but unknown order. *Internat. Econom. Rev.*, 22:55–70, 1981.

[10] X. Guyon. *Random Fields on a Network.* Springer-Verlag, 1995.

[11] X. Guyon and C. Hardouin. The Chi-square coding test for nested Markov random fields hypotheses. L. N. S. 74, pages 165–176. Springer-Verlag, 1992.

[12] E. J. Hannan. The estimation of the order of an ARMA process. *Ann. Statist.*, 8:1071–1081, 1980.

[13] E. J. Hannan. Estimating the dimension of a linear system. *J. Multivariate Anal.*, 11:459–473, 1981.

[14] E. J. Hannan and M. Kanter. Autoregressive processes with infinite variance. *J. Appl. Prob.*, 14:411–415, 1977.

[15] E.J. Hannan and M. Deistler. *The Statistical Theory of Linear Systems.* Wiley, New York, 1988.

[16] W. Hoeffding. Probability inequality for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:12–30, 1963.

[17] Ph. Jolivaldt. *Contrôle de l'ensemble de bon choix de modèle dans un problème d'identification.* PhD thesis, Université Paris I, 1992.

[18] K. Knight. Consistency of Akaike's information criterion for infinite variance autoregressive processes. *Ann. Statist.*, 17:824–840, 1989.

[19] T. Mikosch, T. Gadrich, C. Klüppelberg, and R. Adler. Parameter estimation for ARMA models with infinite variance innovations. *Ann. Statist.*, pages 305–326, 1995.

[20] J. Paulsen. Order determination of multivariate autoregressive time series with unit roots. *J. Time Ser. Anal.*, 5:115–127, 1984.

[21] V. V. Petrov. *Limit Theorems of Probability Theory*. Clarendon Express, Oxford, 1995.

[22] B. M. Pötscher. Model selection under nonstationarity: autoregressive models and stochastic linear regression models. *Ann. Statist.*, 17:1257–1274, 1989.

[23] B. M. Pötscher. Non invertibility and pseudo-maximum likelihood estimation of misspecified ARMA models. *Econometric Theory*, 7:435–449, 1991.

[24] B. G. Quinn. Order determination for a multivariate autoregression. *J. Roy. Statist. Soc. Ser. B*, 42:182–185, 1980.

[25] B. Ripley. *Statistical Inference for Spatial Processes*. Cambridge Univ. Press, 1988.

[26] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6:461–464, 1978.

[27] R. Senoussi. Statistique asymptotique presque-sûre des modèles statistiques convexes. *Ann. Inst. Henri Poincaré*, 26:19–44, 1990.

[28] R. Shibata. Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, 63:117–126, 1976.

[29] R. S. Tsay. Order selection in nonstationary autoregressive models. *Ann. Statist*, 12:1425–1433, 1984.

[30] P. Whittle. On stationary processes in the plane. *Biometrika*, 41:434–449, 1954.

[31] P. Zhang. On the convergence rate of model selection criteria. *Communications in Statist., Theory Meth.*, 22(10):2765–2775, 1993.

[32] L.C. Zhao, P.R. Krishnaiah, and Z.D. Bai. On detection of the number of signals when the noise covariance matrix is arbitrary. *J. Multivariate Anal.*, 20:26–49, 1986.

A Proof of Proposition 5

It is already shown that **(C1)** holds. For **(C2)**, we find for both models

$$U(\theta,\gamma_0) - U(\theta_0,\gamma_0) = (2\pi)^{-d} \int_{\mathbb{T}^d} \left[\frac{f_{\theta_0}}{f_\theta} - 1 - \log \frac{f_{\theta_0}}{f_\theta} \right] (\lambda) d\lambda \geq 0,$$

where the equality holds if and only if $\theta = \theta_0$. Therefore θ_0 is the unique minimum of $U(\theta, \gamma_0)$ on Θ.

We now check the smoothness condition **(C3-i)** and **(D)** separately for the two models.

The AR(m) model. By (34) and (32), we get

$$U(\theta, \gamma) = \log \sigma_\varepsilon^2 + \sigma_\varepsilon^{-2} \left[{}^t\theta \Gamma \theta - 2\, {}^t\theta u + \gamma(0) \right] , \tag{60}$$

where Γ is the $m \times m$ auto-covariance matrix $[\gamma(i-j)]_{1 \le i,j \le m}$, with $\gamma(-j) = \gamma(j)$, and $u = [\gamma(j)]_{1 \le j \le m}$. For each submodel $P \subseteq M$ and $\theta_P = (\theta_i \mathbf{1}_{i \in P})$, the function $\theta_P \mapsto U(\theta_P, \gamma)$ is a positive quadratic map, having a unique minimum $\theta_P(\gamma) = \Gamma_P^{-1} u_P$. Clearly $\gamma \mapsto \theta_P(\gamma)$ is everywhere continuous and **(D)** holds. □

The CAR(m) model. The parameter space Θ given by (30) is convex. Define $g_\theta(\lambda) := \sigma_\varepsilon^2 [(2\pi)^d f_\theta(\lambda)]^{-1} = 1 - 2 \sum_M \theta_\ell \cos\langle \lambda, \ell \rangle$. The function $\theta \mapsto U(\theta, \gamma)$ is convex because its Hessian matrix,

$$D(\theta) = U^{(2)}_{\theta^2}(\theta, \gamma) = 4(2\pi)^{-d} \left(\int_{\mathbb{T}^d} \frac{\cos\langle u, \lambda \rangle \cos\langle v, \lambda \rangle}{g_\theta(\lambda)^2} d\lambda \right)_{u,v \in M} , \tag{61}$$

is positive definite on Θ. Thus any minimum $\theta_P(\gamma)$ of $U(\theta, \gamma)$ on Θ_P, if it exists, will be unique. It remains to prove the existence of such a minimum. For this, we shall show that $U(\theta, \gamma)$ tends to infinity when θ approaches the boundary of Θ_P.

First note that Θ is bounded ($||\theta||_\infty \le 1$ on Θ) : indeed, an application of the Fourier inversion formula to the positive polynomial g_θ yields

$$|-\theta_\ell| = \left| (2\pi)^{-d} \int_{\mathbb{T}^d} g_\theta(\lambda) \exp -i\langle \lambda, \ell \rangle d\lambda \right| \le (2\pi)^{-d} \int_{\mathbb{T}^d} g_\theta(\lambda) d\lambda = 1 \ .$$

Set $G(\theta) = -(2\pi)^{-d} \int_{T^d} \log g_\theta(\lambda) d\lambda$ and $H(\theta, \gamma) = U(\theta, \gamma) - G(\theta)$. As Θ is bounded, H is bounded on Θ.

Let $\overline{\theta}$ be some boundary point of Θ_P. By the definition of Θ, there exists some $\mu^* \in \mathbb{T}^d$ s.t. $g_{\overline{\theta}}(\mu^*) = 0$. As $g_{\overline{\theta}} \ge 0$, we find $g^{(1)}_{\overline{\theta},\lambda}(\mu^*) = 0$. A Taylor expansion at μ^*, together with the compactness of $\mathbb{T}^d$ ensure that there is some $a > 0$ s.t. $g_{\overline{\theta}}(\lambda) \le a||\lambda - \mu^*||_2^2$ for all $\lambda \in \mathbb{T}^d$. It follows that $G(\overline{\theta}) = \infty$. Take some sequence (θ_n) converging to $\overline{\theta}$. Then (g_{θ_n}) converges uniformly to $g_{\overline{\theta}}$. Hence $\lim_{\theta_n \to \overline{\theta}} G(\theta_n) = G(\overline{\theta}) = \infty = \lim_{\theta_n \to \overline{\theta}} U(\theta, \gamma)$. The existence of an (unique) minimum $\theta_P(\gamma) \in \Theta_P$ follows. Thus **(C3-i)** is proved.

Finally by (61), the second order smoothness **(D)** obviously holds. ■

B Proof of Proposition 6

The assumption **(C1)** holds by definition (37) of U. Let us examine **(C2)**. The assumption **(CAT)** ensures that the θ-parametrisation is *proper*, i.e. the map $\theta \mapsto \pi_\theta$ is injective. Denote by $\mathbb{E}_{\theta,v}$ and $\mathbb{V}_{\theta,v}$ the expectation and the variance under $\pi_\theta(\cdot|v)$, respectively. For $\gamma = (\pi_\omega(a|v)) \in F$,

$$U(\theta, \gamma) = -\sum_{v \in V} \frac{n_v}{n} \mathbb{E}_{\omega,v} \left[\log \pi_\theta(X|v) \right] . \tag{62}$$

The r.h.s. of the above equation is a weighted sum of Kullback-Leibler discrepancy between the conditional distributions $\pi_\omega(\cdot|v)$ and $\pi_\theta(\cdot|v)$. Taking $\gamma = \gamma_0 = (\pi_{\theta_0}(a|v))$, we find that θ_0 is the unique minimum of the map $\theta \mapsto U(\theta, \gamma_0)$ on $\Theta = \mathbb{R}^m$. Hence **(C2)** holds.

We now check **(C3-i)** and **(D)**. Fix some $P \subseteq M$. It is easy to see that for all γ, the map $\theta_P \mapsto U(\theta_P, \gamma)$ has continuous second order derivatives on Θ_P (hereafter, we drop the index P in θ). In particular, its Hessian matrix at (θ, γ) is equal to

$$U^{(2)}_{\theta^2}(\theta, \gamma) = \sum_v \frac{n_v}{n} \mathbb{V}_{\theta,v} [\boldsymbol{\phi}(X, v)] \ , \tag{63}$$

which is independent from γ. It will be shown below, while proving **(D),** that this matrix is positive definite. Because Θ_P is convex and $\theta \mapsto U(\theta, \gamma)$ is strictly convex, any minimum $\theta_P(\gamma)$ of $U(\theta, \gamma)$ on Θ_P will be unique (if it exists). It remains to prove the existence of such a minimum. For this, we shall prove that $U(\theta, \gamma) \to \infty$ as $||\theta|| \to \infty$.

Let us take some non null vector D in Θ_P and consider $\theta = \beta D \in \Theta_P$ while letting $\beta \to \infty$. For any $v \in V$, let us define $W_v(a) = \langle D, \boldsymbol{\phi}(a, v) \rangle$, and its maximum $\overline{W}_v = \max\{W_v(a) : a \in E\}$. Thus

$$\mathbb{E}_{\omega,v} \log \pi_\theta(X|v) = \beta \mathbb{E}_{\omega,v} [W_v(X)] - \log \sum_{a \in E} \exp [\beta W_v(a)] \ . \tag{64}$$

When $\beta \to \infty$, we have $\sum_{a \in E} \exp [\beta W_v(a)] = \alpha_v e^{\beta \overline{W}_v} [1 + \varepsilon_v(\beta)]$ with some positive integer α_v and $\varepsilon_v(\beta) \to 0$. Hence by (62) and (64)

$$U(\theta, \gamma) = \beta \sum_{v \in V} \frac{n_v}{n} \mathbb{E}_{\omega,v} \left[\overline{W}_v - W_v(X)\right] + \sum_{v \in V} \frac{n_v}{n} \left[\log \alpha_v [1 + \varepsilon_v(\beta)]\right].$$

On the other hand, the assumption **(CAT)** ensures that there is some $v_* \in V$ for which the map $a \mapsto W_{v_*}(a)$ is not constant. It follows that $U(\theta, \gamma) \to \infty$ as $\beta \to \infty$. The assumption **(C3-i)** is proved.

Finally to check **(D)**, it is enough to prove that the Hessian $U^{(2)}_{\theta^2}(\theta, \gamma)$ in (63) is positive definite. Let us take again some $D \in \mathbb{R}^m \backslash \{\mathbf{0}\}$. With the same W_v's and v_* as defined before, the conclusion follows from

$$D' U^{(2)}_{\theta^2}(\theta, \gamma) D = \sum_{v \in V} \frac{n_v}{n} \mathbb{V}_{\omega,v} W_v(X) \geq \frac{n_{v_*}}{n} \mathbb{V}_{\omega,v_*} W_{v_*}(X) > 0. \quad \blacksquare$$

C Proof of Proposition 8

For a linear process as defined in (50), there is no basic difference between the unidimensional case $d = 1$ and the multidimensional case $d \geq 2$. Therefore we shall hereafter assume that $d = 1$ for simplicity. Since $T_n = [\widehat{\gamma}_n(\ell) : \ell \in M \cup \{0\}]$ and $\mathbb{E}_0 \widehat{\gamma}_n(\ell) = \gamma_0(\ell)$, we have to control deviation probabilities $\mathbb{P}_0(|\widehat{\gamma}_n(\ell) - \gamma_0(\ell)| \geq x)$: the intended result follows from specifications of such estimates with suitable values of x as indicated in Theorem 1. Again for simplification purposes, we shall explain in detail only the case $\ell = 0$. For $\ell \neq 0$, the results follow in a similar way.

Following [2], the idea of the proof is to use the linear representation (50). Since there is some trouble in their proof (see Eqs. (3.23) and (3.27) there), we reconsider it here. For $\ell = 0$,

$$\widehat{\gamma}_n(0) - \sigma_X^2 = A_n + B_n \ , \tag{65}$$

with

$$A_n = \frac{1}{n}\sum_{t=1}^{n}\sum_{s\geq 0} a_s^2(\varepsilon_{t-s}^2 - \sigma_\varepsilon^2), \quad B_n = \frac{2}{n}\sum_{t=1}^{n}\sum_{\Delta>0}\sum_{s\geq 0} a_s a_{s+\Delta}\varepsilon_{t-s}\varepsilon_{t-s-\Delta} \ .$$

Thus for $x > 0$,

$$\mathbb{P}(|\widehat{\gamma}_n(0) - \sigma_X^2| \geq x) \leq \mathbb{P}(|A_n| \geq \frac{x}{2}) + \mathbb{P}(|B_n| \geq \frac{x}{2}) \ .$$

We estimate the right hand side in several steps. Recall that $\alpha \in (0,1)$ is the rate such that $|a_s| \leq A\alpha^s$. The following constants will be used in the sequel.

$$\beta \in (\alpha, 1), \quad h = \beta/\alpha, \quad n_* = \min\{k \geq 0 : h^k \leq k\} \ , \tag{66}$$

$$K_0 = (4\beta A^2)^{-1}(1-\beta)(1-\beta^2), \quad K_0' = \frac{1}{2}A^{-2}(1-\beta^2) \ , \tag{67}$$

$$q = \frac{\tau g(m+1)}{K_0\delta_0^+}, \quad q' = \frac{\tau g(m+1)}{K_0'\delta_0^+} \ , \tag{68}$$

$$\sigma_1 = q^2, \quad \sigma_2 = \left(\frac{q}{h}\right)^2 \wedge (q')^2 \wedge \eta_0^+ \ . \tag{69}$$

Step (1). Estimate for $\mathbb{P}(|A_n| > \frac{1}{2}x)$. Set $E_t := \varepsilon_t^2 - \sigma_\varepsilon^2$, $x_s = K_0' h^{2s} x$, $\beta_s^2 = (1-\beta^2)\beta^{2s}$. Since $\sum_{s\geq 0}\beta_s^2 = 1$,

$$\left\{|A_n| \geq \frac{x}{2}\right\} \subseteq \bigcup_{s\geq 0}\left\{a_s^2\left|\frac{1}{n}\sum_{t=1}^{n} E_{t-s}\right| \geq \frac{1}{2}\beta_s^2 x\right\} \subseteq \bigcup_{s\geq 0}\left\{\left|\frac{1}{n}\sum_{t=1}^{n} E_{t-s}\right| \geq x_s\right\} .$$

E_t has an exponential moment $\mathbf{E}(\tau, g)$ for some positive constants τ and g. Set $\xi_{\tau,g}(x) = \frac{1}{2}\left[(x^2/g) \wedge (\tau x)\right]$. Therefore

$$\mathbb{P}\left\{|A_n| \geq \frac{x}{2}\right\} \leq 2\sum_{s\geq 0} e^{-n\xi_{\tau,g}(x_s)} \ . \tag{70}$$

Step (2). Estimate for $\mathbb{P}(|B_n| > \frac{1}{2}x)$. Similarly, set for any positive integer Δ,

$$F_{t,\Delta} = \varepsilon_t\varepsilon_{t-\Delta}, \qquad \mathbf{F}_{s,\Delta} = n^{-1}\sum_{t=1}^{n} F_{t-s,\Delta} \ ,$$
$$x_{s,\Delta} = K_0 h^{2s+\Delta} x \ , \quad \beta_{s,\Delta} = (1-\beta^2)(1-\beta)\beta^{2s+\Delta-1} \ .$$

Since $\sum_{s\geq 0,\,\Delta\geq 1}\beta_{s,\Delta} = 1$, we may write

$$\left\{|B_n| \geq \frac{x}{2}\right\} \subseteq \bigcup_{s\geq 0,\,\Delta\geq 1}\left\{\left|\frac{1}{n}\sum_{t=1}^{n} F_{t-s,\Delta}\right| \geq x_{s,\Delta}\right\} = \bigcup_{s\geq 0,\,\Delta\geq 1}\{|\mathbf{F}_{s,\Delta}| \geq x_{s,\Delta}\} \ .$$

Hence

$$\mathbb{P}\left\{|B_n| \geq \frac{x}{2}\right\} \leq \sum_{\Delta\geq 1}\sum_{s\geq 0}\mathbb{P}\{|\mathbf{F}_{s,\Delta}| \geq x_{s,\Delta}\} \ . \tag{71}$$

Fix some $s \geq 0$ and define $I_n = \{1, \ldots, n\}$, $J_1 = \{t : t \in I_n \text{ and } 1 \leq [t] \leq \Delta\}$with $[t] := t \mod 2\Delta$, $J_2 = I_n - J_1$, $n_k = |J_k|$ and $\mathbf{F}_k = n_k^{-1}\sum_{t\in J_k} F_{t-s,\Delta}$ for $k = 1, 2$ (with the

convention $\mathbf{F}_k = 0$ if $n_k = 0$). Such a decomposition of I_n ensures the independence of the variables $\{F_{t-s,\Delta} : t \in J_k\}$ in each subset J_k. Furthermore, since $n_1 \geq n_2$, we have that $n_1 \geq \frac{1}{2}n > 0$. Then

$$\{|\mathbf{F}_{s,\Delta}| \geq x_{s,\Delta}\} \subset \bigcup_{k=1,2} \{|\mathbf{F}_k| \geq x_{s,\Delta}\} \ . \tag{72}$$

On the other hand, consider the variable $F = \varepsilon_u \varepsilon_v$ for some $u \neq v$. Since $|F| \leq \frac{1}{2}(\varepsilon_u^2 + \varepsilon_v^2)$, (40) holds for F. Thus F has an exponential moment $\mathbf{E}(\tau', g')$ for some positive constants τ' and g'.

We may assume that $\tau' = \tau$ and $g' = g$ (otherwise replace τ, τ' by $\tau \wedge \tau'$, and g, g' by $g \vee g'$). Then ξ will hereafter denote this common bound $\xi_{\tau,g}$ for the variables $\{\varepsilon_t^2 - \sigma^2\}$ and $\{F_{t,\Delta}\}$.

We now split the r.h.s of (71) in three different terms according to whether $\Delta \geq n$, $\Delta \leq \frac{1}{2}n$ or $\frac{1}{2}n < \Delta < n$. If $\Delta \geq n$, $n_2 = 0$ and $n = n_1$. Hence

$$\mathbb{P}\{|\mathbf{F}_{s,\Delta}| \geq x_{s,\Delta}\} \leq 2e^{-n\xi(x_{s,\Delta})} \ . \tag{73}$$

If $\Delta \leq \frac{1}{2}n$, then $n_1 \geq \frac{1}{3}n$ and $n_2 \geq \frac{1}{3}n$. Therefore

$$\mathbb{P}\{|\mathbf{F}_{s,\Delta}| \geq x_{s,\Delta}\} \leq \sum_{k=1,2} \mathbb{P}\{|\mathbf{F}_k| \geq x_{s,\Delta}\} \leq 4e^{-\frac{1}{3}n\xi(x_{s,\Delta})} \ . \tag{74}$$

In the last case, since $n_2 \geq 1$, $n_1 \geq 1$ and ξ is increasing, we have

$$\mathbb{P}\{|\mathbf{F}_{s,\Delta}| \geq x_{s,\Delta}\} \leq \sum_{k=1,2} \mathbb{P}\{|\mathbf{F}_k| \geq x_{s,\Delta}\} \leq 4e^{-\xi(x_{s,\Delta})} \ . \tag{75}$$

Collecting (71) to (75) yields

$$\frac{1}{4}\,\mathbb{P}\left\{|B_n| \geq \frac{x}{2}\right\} \leq \left(\sum_{\Delta \geq n} + \sum_{\Delta \leq \frac{1}{2}n}\right) \sum_{s \geq 0} e^{-\frac{1}{3}n\xi(x_{s,\Delta})} + \sum_{\frac{1}{2}n < \Delta < n} \sum_{s \geq 0} e^{-\xi(x_{s,\Delta})} \tag{76}$$

Step (3). An auxiliary lemma. The upper bounds in (76) and (70) show that we have to estimate sums of type $\sum_{s>0} \exp\left[-\alpha_n \xi(x_{s,\Delta})\right]$ with some $\alpha_n > 0$. This is done in the following lemma that we shall prove later.

Lemma 13 *(i) Let $p \in (0,1)$, $u > 1$. For any integers $1 \leq s \leq t$,*

$$\sum_{k=s}^{t} p^{u^k} \leq \frac{1}{\log u \,|\log p|} \frac{p^{u^{s-1}}}{u^{s-1}}. \tag{77}$$

In particular $\sum_{s \geq 0} p^{u^s} \leq 1 + (\log u \,|\log p|)^{-1}$.

(ii) By setting $K_1 = \frac{1}{2}g^{-1}K_0^2$, $K_2 = \frac{1}{2}\tau K_0$, $u(x) = \log[\tau g/(K_0 x)]/\log h$ and $v(x) = 1 + (2K_2 x \log h)^{-1}$, we have

$$\sum_{s \geq 0} \exp\left[-\alpha_n \xi(x_{s,\Delta})\right] \leq \left[1 + \frac{1}{2}u(x)\right] \mathbf{1}_{\{u(x) \geq \Delta\}}\, e^{-\alpha_n K_1 x^2 h^{2\Delta}} + v(x)\, e^{-\alpha_n K_2 x h^{\Delta}} \ , \tag{78}$$

(iii) Similarly by setting $K_1' = \frac{1}{2}g^{-1}K_0'^2$, $K_2' = \frac{1}{2}\tau K_0'$, $u'(x) = \log[\tau g/(K_0' x)]/\log h$ *and* $v'(x) = 1 + (2K_2' x \log h)^{-1}$*, we have*

$$\sum_{s\geq 0} \exp\left[-\alpha_n \xi(x_s)\right] \leq \left[1 + \frac{1}{2}u'(x)\right] \mathbf{1}_{\{u'(x)\geq 0\}}\, e^{-\alpha_n K_1' x^2} + v'(x)\, e^{-\alpha_n K_2' x}\,. \tag{79}$$

Step (4). Final estimate for $\mathbb{P}(|\widehat{\gamma}_n(0) - \sigma_X^2| \geq x)$. First an application of (79) with $\alpha_n = n$ to the r.h.s of (70) yields

$$\frac{1}{2}\mathbb{P}\left\{|A_n| \geq \frac{x}{2}\right\} \leq \left[1 + \frac{1}{2}u'(x)\right] \mathbf{1}_{\{u'(x)\geq 0\}}\, e^{-nK_1' x^2} + v'(x)\, e^{-nK_2' x} =: E_1 + E_2\ . \tag{80}$$

Let $\widetilde{u}(x) := u(x)\left[1 + \frac{1}{2}u(x)\right]$. For the first sum in the r.h.s (76), we find by successive applications of (77)-(78)

$$\begin{aligned}
&\left(\sum_{\Delta\geq n} + \sum_{\Delta\leq \frac{1}{2}n}\right) e^{-\frac{1}{3}n\xi(x_{s,\Delta})} \\
&\leq \left(\sum_{\Delta\geq n} + \sum_{\Delta\leq \frac{1}{2}n}\right)\left(\left[1 + \frac{1}{2}u(x)\right] \mathbf{1}_{\{u(x)\geq\Delta\}}\, e^{-\frac{1}{3}nK_1 x^2 h^{2\Delta}} + v(x)\, e^{-\frac{1}{3}nK_2 x h^{\Delta}}\right) \\
&\leq \left(\sum_{\Delta\geq n} + \sum_{\Delta\leq \frac{1}{2}n}\right)\left[1 + \frac{1}{2}u(x)\right] \mathbf{1}_{\{u(x)\geq\Delta\}}\, e^{-\frac{1}{3}nK_1 x^2 h^{2\Delta}} + \sum_{\Delta\geq 1} v(x)\, e^{-\frac{1}{3}nK_2 x h^{\Delta}} \\
&\leq \mathbf{1}_{\{u(x)\geq n\}}\tilde{u}(x) e^{-\frac{1}{3}nK_1 x^2 h^{2n}} + \mathbf{1}_{\{u(x)\geq 1\}}\tilde{u}(x) e^{-\frac{1}{3}nK_1 x^2 h^2} + \frac{3v(x)}{nK_2 x\log h} e^{-\frac{1}{3}nK_2 x} \\
&=: E_3 + E_4 + E_5\ .
\end{aligned} \tag{81}$$

Similarly, for the last sum in (76),

$$\begin{aligned}
\sum_{\frac{1}{2}n<\Delta<n}\sum_{s\geq 0} e^{-\xi(x_{s,\Delta})} &\leq \sum_{\frac{1}{2}n<\Delta<n}\left(\left[1 + \frac{1}{2}u(x)\right] \mathbf{1}_{\{u(x)\geq\Delta\}}\, e^{-K_1 x^2 h^{2\Delta}} + v(x)\, e^{-K_2 x h^{\Delta}}\right) \\
&\leq \mathbf{1}_{\{u(x)>\frac{1}{2}n\}}\tilde{u}(x) e^{-\frac{1}{3}nK_1 x^2 h^{2n}} + \frac{v(x)}{K_2 x(\log h) h^{\frac{n}{2}-1}} e^{-K_2 x h^{\frac{n}{2}-1}} \qquad (82) \\
&=: E_6 + E_7 \qquad (83)
\end{aligned}$$

Collecting all these estimates gives

$$\mathbb{P}(|\widehat{\gamma}_n(0) - \sigma_X^2| \geq x) \leq E_1 + \cdots + E_7. \tag{84}$$

Estimate for $\mathbb{P}(M_n^-)$. Recall that under the assumptions, we have $c_n/n \leq \eta_0^-$ and $n \geq n_*$ (hence $h^{n/2} \geq n$). Since $M_n^- \subseteq \{||T_n - \gamma_0|| \geq \delta_0^-\} \subseteq \bigcup_{\ell=0}^m \{|\widehat{\gamma}_n(\ell) - \gamma_0(\ell)| \geq (m+1)^{-1}\delta_0^-\}$,

we may apply (84) with $x = (m+1)^{-1}\delta_0^-$. Taking into account the condition $h^{n/2} \geq n$, one easily checks that $\mathbb{P}(|\widehat{\gamma}_n(0) - \sigma_X^2| \geq (m+1)^{-1}\delta_0^-) \leq \mu_0 e^{-n\mu_0'}$ for some constants $\mu_0 \geq 0$ and $\mu_0' > 0$. As we also have $\mathbb{P}(|\widehat{\gamma}_n(\ell) - \gamma_0(\ell)| \geq (m+1)^{-1}\delta_0^-) \leq \mu_\ell e^{-n\mu_\ell'}$ for some constants $\mu_\ell \geq 0$ and $\mu_\ell' > 0$, $\ell = 1, \ldots, m$, we take $\mu = \mu_0 + \cdots + \mu_m$ and $\mu' = \min\{\mu_0', \ldots, \mu_m'\}$ to conclude the first part (i) of the proposition.

Estimate for $\mathbb{P}(M_n^+)$. Here we apply (84) with $x = (m+1)^{-1}\delta_0^+\sqrt{\frac{c_n}{n}}$. Under the assumption $nc_n \geq q^2$, it is readily checked that $n \geq 2u(x)$. Therefore, $E_3 = E_6 = 0$. Taking into account the conditions $n \geq n_*$ (hence $h^{n/2} \geq n$) and $c_n/n \leq (q_1/h)^2 \wedge q_2^2 \wedge \eta_0^+$, we find

$$\begin{aligned}
&\mathbb{P}\left\{|\widehat{\gamma}_n(0) - \sigma_X^2| \geq (m+1)^{-1}\delta_0^+\sqrt{\frac{c_n}{n}}\right\} \\
&\quad\leq \quad \nu_1\left\{1 + \log\frac{n}{c_n} + \left(\log\frac{n}{c_n}\right)^2\right\} e^{-\nu_1' c_n} + \nu_2\left\{1 + \sqrt{\frac{n}{c_n}} + \frac{1}{c_n}\right\} e^{-\nu_2'\sqrt{nc_n}}
\end{aligned}$$

with some constants $\nu_j \geq 0$ and $\nu_j' > 0$ ($j = 1, 2$). As for (i), the result (ii) of the proposition follows by summing these inequalities over $\ell = 0, \ldots, m$. ■

To complete the proof of Proposition 8, it remains to prove Lemma 13.

Proof of Lemma 13 . Part (i) follows from the elementary inequality $\sum_{k=s}^t p^{u^k} \leq \int_{s-1}^t p^{u^x} dx$.

Part (ii). As $\xi(x) = \frac{x^2}{2g} \wedge \frac{\tau|x|}{2}$, we have $\xi(x) = \frac{x^2}{2g}$ if $|x| \leq \tau g$, and $\xi(x) = \frac{1}{2}\tau|x|$ otherwise. Since $x_{s,\Delta} = K_0 h^{2s+\Delta}x$, one has $|x_{s,\Delta}| \leq \tau g$ if and only if $s \leq s^* := \frac{1}{2}[u(x) - \Delta]$. Assume that $u(x) \geq \Delta$. For $s \leq s^*$, $\exp[-\alpha_n\xi(x_{s,\Delta})] \leq \exp[-\alpha_n\xi(x_{0,\Delta})] = \exp[-\alpha_n K_1 x^2 h^{2\Delta}]$. As $(1 + s^*) \leq \left[1 + \frac{1}{2}u(x)\right]$, we find

$$\sum_{s \leq s*} \exp\left[-\alpha_n\xi(x_{s,\Delta})\right] \leq \left[1 + \frac{1}{2}u(x)\right] \mathbf{1}_{\{u(x)\geq\Delta\}}\, e^{-\alpha_n K_1 x^2 h^{2\Delta}}.$$

For $s > s^*$, we have that $\xi(x_{s,\Delta}) = K_2 x h^{2s} h^\Delta$. By applying Part (i) with $p = \exp[-\alpha_n K_2 x h^\Delta]$ and $u = h^2$, it follows that

$$\sum_{s > s^*} \exp\left[-\alpha_n\xi(x_{s,\Delta})\right] \leq \sum_{s\geq 0} \exp\left[-\alpha_n K_2 x h^{2s} h^\Delta\right] \leq v(x)\, e^{-\alpha_n K_2 x h^\Delta}\ .$$

Part (iii) follows in the same way as for Part (ii). ■